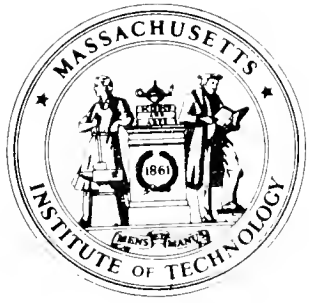

A NOTE ON REGRESSION ANALYSIS AND ITS MISINTERPRETATIONS

Peer Soelberg

247-67

A NOTE ON REGRESSION ANALYSIS AND ITS MISINTERPRETATIONS

Peer Soelberg

247-67

This note, which collegues have urged be put into wider circulation, was first written as a warning to participants in the author's Doctoral Seminar on Research Methods not to interpret computerized regression analyses invalidly.

REGRESSION ANALYSIS AND ITS MISINTERPRETATIONS

Least square multivariate regression analysis -- such as gets carried out automatically by computer programs like the Soelberg "Adaptive Muliple Regression Analysis"* or the Beaton and Glauber "Statistical Laboratory Ultimate Regression Package" (SLURP)** -- is strictly applicable only under the following set of assumptions:

A. Functional Form

The functional form of the relationship between y (the dependent variable) and X (the independant variables) is known and specified a priori. Linear relationships are commonly assumed, but any functional form may be specified, provided that the other assumptions (below) are not thereby violated.

B. Complete Specifications of Variables

The vector X of independent variables includes all variables that exert a systematic effect on y; or the systematic effects on y of non-included not-X exactly counterbalance each other; or the not-X which do exert a systematic effect on y were held constant when the data were collected, in which case the regression prediction must be qualified by the ceteris paribus qualification: "provided that the systematic not-X take on whatever values they had when the estimation data were collected" (or residually affect y in a no different manner).

---

* P. Soelberg, "Adaptive Multiple Regression Analysis", Behavioral Theory of the Firm Paper No. 35, Carnegie Institute of Technology, 1961, 35 pp. plus G-20 GATE program listing.

** A.E. Beaton and R.R. Glauber, Statistical Laboratory Ultimate Regression Package, Harvard Statistical Laboratory, 1962, 33 pp.

Assumption B is usually expressed statistically as the requirement that the expected difference between observed y_i and predicted $\hat{y}_i$ must be zero; i.e.

$$\mathcal{E}(u) = 0$$

where u_i is the error in y_i due, say, to unsystematically faulty measurements of y and/or "random" effects of the uncontrolled not-X.

C. Nature of Independent Variables

The matrix of observation, X, is a set of fixed numbers, i.e. is made up of the same set of X vectors for each new sample or experiment, where the X are not subject to measurement or classification error. However, should the X be subject to the latter type of error, say they were drawn randomly from a population of X, parameters of the regression equation may still be derived (by formulae that are only slightly different from the fixed-X case, but which give wider standard errors of estimate) provided that the errors in X are normally distributed, and that σ_X are known, or can be estimated, a priori.

D. Homoscedasticity

The error of estimate in y remains constant over the whole range of encountered values of the X. If, however, hetroscedesticity is known to exist the data may be transformed accordingly, before normal regression formulae are applied. The estimators will then be unbiased (meaning that they remain maximum likelihood estimates of the true values), but are less efficient than the homoscedastic ones (meaning that a larger sample of observations is required in order to obtain as small an error of estimate, i.e., as narrow a "band" of confidence limits on the sample estimates).

E. Independence of Error

The error of estimate y_i in each y_i observation is uncorrelated with the error y_j in any other observation y_j; specifically there is no serial correlation between one observation and the next.

Serial correlation, i.e. systematic bias in the estimate $\hat{y}$ away from the true y, could occur if the form of the regression equation was not properly specified (say a logarithmic relationship exists in fact, whereas the assumed regression equation tried to minimize differences from a straight line). Another reason for serial correlation might be a systematic biasing effect of uncontrolled not-X variables, which could conceivably move $\hat{y}$ away from the true y in unidentified phases. A moving average, for example, if such were used as an independent variable, is sure to be serially correlated -- obviously, since each new moving average observation is in large part made up of the previous observations.

However, even successive, non over-lapping averages of a series of random numbers will show strong serial correlation patterns.(1) Holbrock Working estimates this serial correlation to be $(m^2-1)/2(2m^2+1)$, where m is the number of elements in the average. Thus it has been argued that if individual stock price changes indeed were a random chain (they apparently exhibit little serial correlation if viewed as a population of separate elements), their averages could still exhibit regular business-cycle type patterns. A similar argument holds, of course, for averages (or volatility differences) of High versus Low prices in a given time period.

If strong autocorrelation exists in $\underline{X}$ the least squares regression estimators will still be unbiased, but the standard error of estimate and the sample variance of the regression coefficients will be seriously underestimated,[2] i.e. standard t-test and F-test tables are no longer valid.

Diagnosis:

A test for the presence of autocorrelated disturbances is available in form of the Durbin Watson d-statistic:[3]

$$d = \sum_{i=2}^{n} ((y_i - y_i) - (y_{i-1} - y_{i-1}))^2 / \sum_{i=1}^{n} (y_i - y_i); \; i = 1,2,....,n \; .$$

Adjusted for the number of explanatory variables, tables are available for critical upper (significant autocorrelation) and lower (no significant autocorrelation) bounds of d. However, for intermediate values of d the Durbin Watson test is inconclusive, i.e. can neither accept nor reject a hypothesis of independence among the regression disturbances.

Cure:

If autocorrelation is rampant, there are corrective methods available. For example, one could assume that the autoregressive scheme was of first order:

$$u_t = \gamma u_{t-1} + \varepsilon$$

where ε are ramdomly distributed with mean and covariance equal to zero,

estimate γ by simple least squares regression, and then transform the original data by the estimate g:

$$y_t^* = y_t - gy_{t-1}; \underline{X}_t^* = \underline{X}_t - g\underline{X}_{t-1} .$$

Durbin has proposed a two-stage estimation procedure which in addition takes into account the effects of higher order autocorrelation (dependences of more than one step) provided that certain regularities in the pattern of dependencies can be assumed.(4)

F. Orthogonality of Independent Variables

The independent variables $\underline{X}$ are assumed to be uncorrelated. If two or more $\underline{X}$ are correlated we run up against the infamous problem of "multicollinearity". Statistical discussion of the nature of this problem are sufficiently confusing, or so shrouded in mathematical lingoism, as to justify the following non-rigorous, intuitive explanation:

The simplest case maybe illustrated for a relationship in three dimensions. Let us assume that the following exact relationship in fact existed among y, X_1, and X_2 :

$$y = X_1/2 + 2X_2$$

Let us also assume that X_1 and X_2 range only from 1 to 2, and take on integer values only. This gives four possible observation vectors or experimental treatments:

X_1 \ X_2	1	2
1	A = 2 1/2	B = 4 1/2
2	D = 2	C = 5

The picture of the regression surface (a two-dimensional plane) would then be the following:

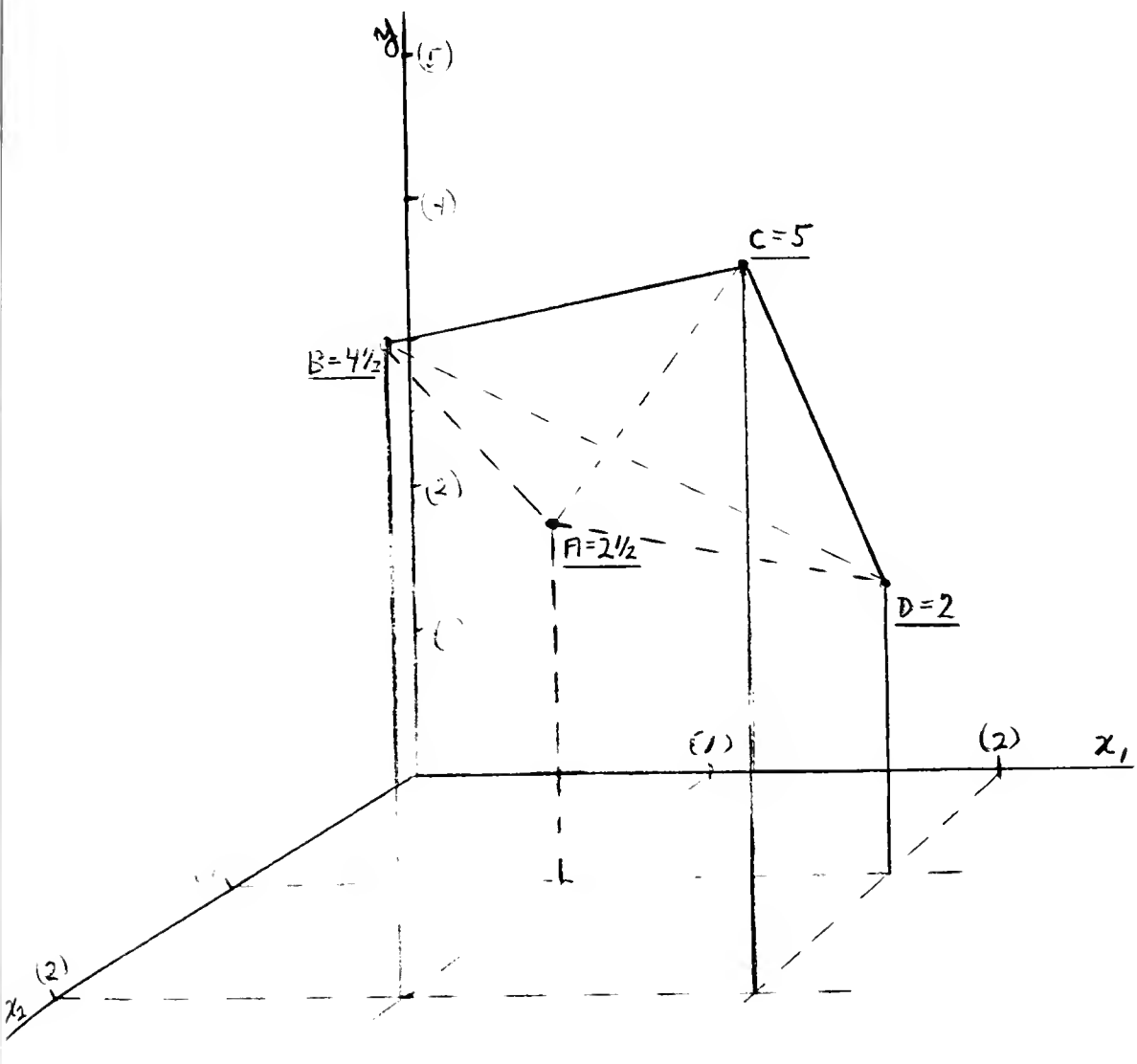

Figure 1

(the regression surface is ABCD)

If X_1 and X_2 were uncorrelated, this simply means that we have obtained an approx. distribution equal of observations in all possible combinations of X cells (i.e. have performed a complete data-collection "experiment"). For example, given 16 observations, in the uncorrelated case these would be uniformly distributed over all possible X values, say thus:

X_1 \ X_2	1	2
1	y_1 y_5 y_9 y_{13}	y_2 y_6 y_{10} y_{14}
2	y_3 y_7 y_{11} y_{15}	y_4 y_8 y_{12} y_{16}

Figure 2

Whereas if X_1 and X_2 were correlated we might conceivably have:

X_1 \ X_2	1	2
1	y_1 y_2 y_5 y_6 y_9 y_{10} y_{13} y_{14}	
2		y_3 y_4 y_7 y_8 y_{11} y_{12} y_{15} y_{16}

Figure 3

(or conversely along the $X_1 = 2$; $X_2 = 2$ diagonal)

The dilemma of multicollinearity is thus clear: In the limit a relatedness among k independent variables provides us with merely enough information to estimate a (n-k) dimensional hyperplane. In our 2-dimensional example (Figure 3) the data would only be sufficient to allow us to estimate the line AC in Figure 1. We would possess no estimate of the angle in which the true ABCD regression plane pivots about that line. Thus we see how in cases of serious, though not perfect, multicollinearity our estimate of the k dimensional "angle" of the regression plane will be determined by the very few observations that happen to fall in the missing X cells, how the estimate will be highly sensitive to any error of estimate deriving from the latter observations.

In general, the standard output of a regression analysis in which any pair or more X are highly collinear cannot be trusted. High R—squares in such cases may be meaningless. The resulting regression coefficients will vary widely whenever one of the "outlying" critical observations is added to or subtracted from the data sample. Standard errors of estimate, or the variances of regression coefficients, will therefore not be much better than nonsense-numbers.

Farrar and Glauber provide a striking illustration of the problem of multicollinearity in their recomputation of the classical economic Cobb Douglas (production function) estimates. The general form of the Cobb Douglas function is assumed to be:

$$P = L^{B1} C^{B2} e^{(\alpha+u)}$$

where P is production, L labor input, and C capital output. Thus:

$$\ln(P) = \alpha + B_1 \ln(L) + B_2 \ln(C) + u$$

to which might be added an additional trend estimator $B_3 t$. Farrar and Glauber then examine six, theoretically trivially different, methods of computing these regression coefficients employing identically the same set of multicollinear data; and include three additional sets of estimates (using the third of the six methods) with <u>one</u> different data point removed from the original set of observations (n = 24):

		B_1	B_2	B_3	R^2
Method	1	.75 (18.1)	$1-B_2$ = .25		.93
	2	.89 (6.0)	$1-B_2$ = .11	.01 (1.0)	.93
	3	.81 (5.6)	.23 (3.7)		.95
	4	.91 (6.5)	-.53 (1.5)	.05 (2.3)	.96
	5	.05	.60		.88
	6	1.35	.01		.92
One data point excluded:	7	.75 (5.0)	.25 (4.0)		.95
	8	.60 (3.3)	.34 (4.0)		.96
	9	1.02 (8.0)	.12 (2.0)		.98

The numbers in parantheses are the coefficients' adjusted t-values. The instability of the coefficients (which have been accepted as gospel by economics texts) should be apparent: B_1 ranges from .05 to 1.35; B_2 from -.53 to .60.

Diagnosis:

In the SLURP program Farrar and Glauber have automated a set of diagnostics of multipcollinearity,[6] of which we will here outline only the interpretation, without its mathematical derivation. Diagnosis of multicollinearity proceeds at three levels:

i. Determination of departure from internal orthogononality in the $\underline{X}$: The determinant of the intercorrelations of $\underline{X}$, i.e. $\underline{X}'\underline{X}$, varies between "zero" (perfect multicollinearity) and "one" (complete orthogonality), over which range this determinant (assuming random draws of sample correlation matrices, and multivariate normal distributions of $\underline{X}$) is distributed approximately like chi-square. Thus to test the hypothesis that there is no more multicollinearity in our sample than what may be expected by chance, we examine the probability for the SLURP reported chi-square value for the $\underline{X}'\underline{X}$ determinant, for the appropriate (n(n-1)/2) degrees of freedom: If this probability is too low (i.e. chi-square too large), we cannot ignore the multicollinearity problem.

ii. Determining which $\underline{X}$ are collinear:
By examining the multiple correlation coefficient between each X_i and the other $\underline{X}$ we can identify which X_i is more or less collinear with the remainer set. As these squared multiple correlation coefficients are distributed approximately like F, SLURP also outputs their F-values, with appropriate degrees of freedom, with which (for appropriately large values) we may have to reject the hypothesis that one or more X_i is not collinear with the rest of the $\underline{X}$ set.

iii. Determining the pattern of interdependence among the collinear X: The pattern of interdependence among the collinear X in the regression equation may be inferred by examining their partial correlation coefficients, to what extent each one is related to another when the effects of all the other X have been "partialled out" of that relationship. (One may view the partial correlation coefficient between X_1 and X_2, corrected for a possible relationship of X_3 to X_1 and of X_3 to X_2, as being the average of the correlation that exist between X_1 and X_2 at each level or value of X_3. Another way of viewing the partial correlation coefficient, say $r^2_{12.3}$, is as the proportion of the variation in X_1 that is left unaccounted for by the relationship of X_3 to X_1 which get explained by the variation in X_2.)

SLURP outputs all these partial correlation coefficients between the X, together with their associated t-values for testing the hypothesis that these partial correlations could have arisen by chance.

Cure

There is no universal cure for multi-collinarity in regression variables. The more tempting (and usually fallacious) method is simply to eliminate all independent variables but one from each collinear set: Granted that this approach may superficially solve the collinearity problem, but by sacrificing information which will be necessary for making valid predictions, provided that the structure of the model, as initially specified, was correctly assumed. In other words, reducing the structural dimensionality of the model will yield predictions that no longer take into account known systematic causes of variation in y (see assumptions A and C above), making the output of a reduced regression analysis potentially highly misleading.

A second method for resolving multicollinearity consists of imposing artificial orthogonality on the independent variables by suitable "rotation" of the dimensions, i.e. by factor analysis, which then might yield a set of more nearly orthogonal set of factors, constructed from linear combinations of the original variables. Regression analysis can then of course be more reliably run on these redefined factors. The problems with this approach (to resolving the multicollinearity dilemma) are identical with the problems of factor analysis per se: i. the weights assumed by the factor loadings over the ranges of the various independent variables will in general not be linear in fact; and ii. the theoretical interpretation of the computationally derived factors will usually be highly unoperational, i.e. no operational definitions or direct measures will in general be available for empirically interpeting the derived factors. Thus the regression coefficients that are derived from analysis of the orthogonal factors will usually have to be retransformed into estimated coefficients for the original X, in which case, unfortunately, all the problems of the original multicollinearity creep right back again into the regression estimates.(7)

The only safe way to cure multicollinearity (if indeed it is curable) derives from the definition of the phenomenon that we presented above, namely to acquire additional information (observations of y) in the non-represented, or sparcely sampled, cells of X-combinations. However, the latter may not occur frequently enough, if at all, in natural experiments--in which case nature may have to be augmented by controlled experimentation, i.e. manipulations, in order to yield the requisite data.

It can be argued that at times some of the X are in fact process-related, such that they will never vary independently of one another -- in which case the general linear hypothesis, i.e. regression models, is not strictly an appropriate mode of analysis (unless the co-varying X are viewed simply as alternate, but noisy, measures of the same underlying variable, in which case "multicollinearity" has arisen from an initially sloppy model specification).

G. Static relationships

The only dynamic relationships that standard regression techniques are equipped to handle in general terms are first-order difference relations with constant coefficients -- and even with these simple equations one often runs into serious estimation problems.(8) In other words, if there is a feedback component in the data one is trying to fit to a regression model the appropriate estimators are likely to be analytically intractable.

In conclusion, we note that whereas slight violation in any one of the above assumptions may be tolerated, depending on the purpose of one's regression analysis, a "reasonable" violation of two or more assumptions can cumulate more than additively. For example, in Monte Carlo studies of mechanisms containing both lagged variables (dynamics) and autocorrelation in residuals Cochrane and Orcutt found that regression analysis produced stable coefficients of more than twice the true size (with standard errors of less than 3%), whereas any one of the violations considered alone would be expected to yield no bias for autocorrelation, and negative bias (underestimation) for lagged variables.(9)

NOTES AND REFERENCES

1. H. Working, "Note on the Correlation of First Differences of Average in a Random Chain", Econometrica, 1960, 28, pp. 916-18

2. For some dramatic examples see the sampling experiments run by D. Cochrane and G. H. Orcutt, "Application of Least Squares Regression to Relationships Containing Auto-Correlated Terms", Journal of the American Statistical Association, 1949, 44, pp. 32-61

3. J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression", I and II, Biometrica, 1950, and 1951; see also H. Theil and A. L. Nager, "Testing the Independence of Regression Disturbances", Journal of the American Statistical Association, 1961, 56, pp. 793-806

4. J. Durbin, "Estimation of Parameters in Time-Series Regression Models", Journal of Royal Statistical Society, 1960, B-22, pp. 139-153

5. D. E. Farrar and R. R. Glauber, "Multi-collinearity in Regression Analysis, The Problem Revisited", M.I.T. and Harvard University, 1964, pp. 7-10

6. Ibid., pp. 30-39

7. Ibid., p.20

8. J. Johnston, Ecnometric Methods, New York: Wiley, 1963, pp. 211-220

9. D. Cochrane and G. H. Orcutt, "A Sampling Study of the Merits of Autoregressive and Reduced Form Transformations in Regression Analysis," Journal of the American Statistical Association, 1949, 44, pp.356-372

DEC 5, 1967

FEB 21 '71

DEC 10 '73

MAY 7 '74